Integration of Data Uncertainty in Linear Regression and Process Optimization

Marco S. Reis and Pedro M. Saraiva
GEPSI-PSE Group, Dept. of Chemical Engineering, University of Coimbra, Pólo II—Pinhal de Marrocos, 3030-290 Coimbra, Portugal

DOI 10.1002/aic.10540
Published online August 9, 2005 in Wiley InterScience (www.interscience.wiley.com).

Data uncertainties provide important information that should be taken into account along with the actual data. In fact, with the development of measurement instrumentation methods and metrology, one is very often able to rigorously specify the uncertainty associated with each measured value. The use of this piece of information, together with raw measurements, should—in principle—lead to more sound ways of performing data analysis, empirical modeling, and subsequent decision making. In this paper, we address the issues of using data uncertainty in the task of model estimation and, when it is already available, we show how the integration of measurement and actuation uncertainty can be achieved in the context of process optimization. Within the scope of the first task (model estimation), we make reference to several methods designed to take into account data uncertainties in linear multivariate regression (multivariate least squares, maximum likelihood principal component regression), and others whose potential to deal with noisy data is well known (partial least squares, principal component regression, and ridge regression), as well as modifications of previous methods that we developed, and compare their performance. MLPCR2 tends to achieve better predictive performance than all the other tested methods. The potential benefits of including measurement and actuation uncertainties in process optimization are also illustrated. © 2005 American Institute of Chemical Engineers *AIChE J,* 51: 3007–3019, 2005

Keywords: measurement uncertainty, multivariate least squares, maximum likelihood principal component regression, partial least squares, principal component regression, optimization under uncertainty

Introduction

The large amounts of industrial and laboratorial data generated in the chemical process industries and stored in databases do have a substantial potential to set the ground for further process improvement and optimization. Noting that this potential is not always being fully developed and that the goals that were present at the conception of such databases are often not being achieved, numerous efforts have been made and documented in the literature toward a more effective use of these information resources, that is, in the fields of process monitoring,[1,2] fault detection and diagnosis,[3] and data mining.[4] However, quite often these approaches do not explicitly and quantitatively take into account data quality, or do so only in an implicit or tacit way. Following the efforts undertaken in the metrology field, with respect to the characterization and quantification of measurement uncertainty, in a rigorous and normalized approach,[5] we believe it is quite appropriate and timely to develop and apply methods that explicitly and consistently take into account this important piece of information.

Measurement uncertainty is a well-defined quantity and there are well-documented standardized procedures that assist its specification or estimation. Basically, uncertainty is defined as a "parameter associated with the result of a measurement

Correspondence concerning this article should be addressed to M. S. Reis at marco@eq.uc.pt.

that characterizes the dispersion of the values that could reasonably be attributed to the measurand." [5] The standard uncertainty u (to which we will often refer simply as "uncertainty") should be expressed in terms of a standard deviation of the values obtained under the same experimental conditions, and can be obtained either from the analysis of collected data (the so-called Type A evaluation) or through other adequate means (Type B evaluation). The availability of raw values, along with their associated uncertainties, implies that we should not have only one data table available for analysis, but in fact two (one with raw values and another one with the corresponding measurement uncertainties). Therefore, with this additional information at our disposal, we should be able to take advantage of it through its integration into our data analysis tasks. For instance, data reconciliation[6-8] is designed to handle noisy measurements, to adjust raw data in some optimal way, so that it conforms to conservation laws and other constraints. The fact that the objective function to be minimized consists of quadratic terms involving the inverse of variance–covariance matrices of measurements[7] indicates that uncertainty information is in fact being considered in data reconciliation. However, there are many application scenarios where no conservation laws are available to perform preliminary data reconciliation, such as the analysis of spectra, microarray data, and laboratorial data sets. Furthermore, uncertainty-based methods can be applied to data sets after reconciliation or filtering.

Sometimes it happens that uncertainty associated with measurements is sufficiently small for the techniques that disregard it completely or treat it in a very simplified way (such as assuming homoscedastic behavior), still holding as adequate. However, these are tacit assumptions, quite often not verified or clearly stated. The main purpose of this article is to bring the issue of data uncertainty into the priorities for the data analyst, which should explicitly address it in a preliminary phase, as well as to present, develop, and test procedures that do exploit and take advantage of data uncertainty information.

In particular, we will address the use of uncertainty information in two different tasks: model estimation and process optimization. In the next section, we refer several methodologies with the potential of integrating uncertainty in the estimation of parameters for a multivariate linear model. This type of model is widely used in the analysis of industrial data sets, and its prediction ability, when parameters are estimated by different methodologies, is thus an important issue in practical applications. In the following section, a complementary situation regarding the use of uncertainties, that is, when a model is considered to be known, is illustrated under the context of process optimization. Then, in the fourth section, we present two case studies that provide the ground for comparison among all the methods referenced in the second section and another case study that illustrates the methodology presented in the third section. We end this paper with a discussion section, where some computational issues are addressed (fifth section) and some final conclusions are drawn. Apart from the comparative study undertaken in the case study section, new methods (unc-PLS3, unc-PLS4, and unc-PLS5) are also presented and tested. Methods MLMLS, unc-PLS1, unc-PLS2, MLPCR2, rMLS, and rMLMLS are carefully described elsewhere.[9] The formulations presented in the third section provide also a contribution to the explicit consideration of measurement uncertainties for process optimization.

Table 1. Formulation of the Optimization Problems Underlying OLS and MLS Methods

OLS	$\hat{b}_{\mathrm{OLS}} = \underset{b=[b_0 \cdots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} [y(i) - \hat{y}(i)]^2 \right\}$	(1)
MLS	$\hat{b}_{\mathrm{MLS}} = \underset{b=[b_0 \cdots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{[y(i) - \hat{y}(i)]^2}{s_e^2(i)} \right\}$	(2)
MLMLS	$\hat{b}_{\mathrm{MLMLS}} = \underset{b=[b_0 \cdots b_p]^T}{\arg\max}\ \Lambda(b)$ $\Lambda(b) = -\frac{1}{2} n \ln(2\pi) - \sum_{i=1}^{n} \ln(\sigma_{\varepsilon_i}) - \frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{[y(i) - \hat{y}(i)]^2}{\sigma_{\varepsilon_i}^2} \right\}$	(3)

Measurement Uncertainties in Model Estimation

This section is devoted to the description of four groups of multivariate linear regression methods that have the potential to accommodate measurement noise information, either explicitly or implicitly. As already mentioned, our focus on multivariate linear regression methods arises from the quite widespread use of this type of approaches in the development of input/output models for industrial and/or laboratorial applications. The several methodologies here addressed are combined under four separate groups, according to their affinity: ordinary least squares (OLS), ridge regression (RR), principal component regression (PCR), and partial least squares (PLS, also referred to as "projection to latent structures"). These four basic methods do not explicitly incorporate measurement uncertainty information, so that several alternatives already developed are also presented, as well as other recent modifications that we propose here and do take uncertainty information explicitly into consideration.

OLS group

Ordinary least squares (OLS) and multivariate least squares (MLS)[10,11] parameter estimates for a linear regression model are the solutions of the optimization problems formulated in Eqs. 1 and 2 of Table 1.

OLS tacitly assumes a homoscedastic behavior (that is, with constant variance) for the noise error term in the standard linear regression model. On the other hand, MLS is built on an error in variables (EIV) functional relationship relating true values of both the input and output variables, which are then affected by zero mean random errors with a given covariance structure (presumed to be known). In the denominator of Eq. 2 we can find a term, $s_e^2(i)$, that results from the summation of the uncertainties associated with the response to those arising from the propagation of uncertainties of the predictors to the response, according to a formula derived from error propagation theory[10,12]:

Table 2. Formulation of the Optimization Problems Underlying RR, rMLS, and rMLMLS

RR	$\hat{b}_{\mathrm{RR}} = \underset{b=[b_0 \cdots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} (y(i) - \hat{y}(i))^2 + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$	(5)
rMLS	$\hat{b}_{\mathrm{rMLS}} = \underset{b=[b_0 \cdots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{(y(i) - \hat{y}(i))^2}{s_e^2(i)} + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$	(6)
rMLMLS	$\hat{b}_{\mathrm{rMLMLS}} = \underset{b=[b_0 \cdots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \ln(s_e(i)) + \sum_{i=1}^{n} \frac{(y(i) - \hat{y}(i))^2}{s_e^2(i)} + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$	(7)

$$s_e^2(i) = uy(i)^2 + \sum_{j=1}^{p} \hat{b}_j^2 uX(i,j)^2 + 2 \sum_{j=2}^{p} \sum_{k=j+1}^{p} \hat{b}_j \hat{b}_k \mathrm{cov}[\Delta\xi_j(i), \Delta\xi_k(i)] \quad (4)$$

where $uX(i, j)$ and $uy(i)$ are the uncertainties associated with the ith observation of the jth input and output variables, respectively, and $\Delta\xi_j(i)$ is the random error affecting the ith measurement of variable j; $\hat{b}_j$ represents the coefficient of the linear regression model associated with variable j.

The method whose objective function is presented in Table 1, Eq. 3, is derived from the analysis of the Berkson case (controlled regressors with error) within the scope of EIV models[13,14] and under the assumption of Gaussian errors. The objective function arises from the maximization of the resulting likelihood function, and we included this approach in our present study given both the similarity between the quadratic functional part of its objective function and the one underlying MLS, and its simplicity. Because the solution for the Berkson case formulation is sometimes similar to MLS,[14] we make reference to the above formulation maximum likelihood multivariate least squares (MLMLS), to stress the statistical origin of the underlying objective function.

RR group

A well-known characteristic of the OLS method is the fact that the variance of its parameter estimates increases when the input variables become more correlated. Computational simulations showed us that the same applies to MLS. One possible way to address this issue consists of enforcing an effective shrinkage in the coefficients under estimation, following a ridge regression (RR) regularization approach. It basically consists of adding an extra term to the objective function that penalizes large solutions (in a square norm sense). Optimization formulations underlying RR estimates,[15,16] as well as those proposed for its counterparts based on MLS and MLMLS, rMLS and rMLMLS, respectively (standing for "ridge MLS" and "ridge MLMLS"), are presented in Table 2.

PCR group

PCR[17,18] is another methodology that handles collinearity among predictor variables. It uses those uncorrelated linear combinations of the input variables that most explain input space variability [from principal components analysis (PCA)] as the new set of predictors, where the response is to be regressed onto. These predictors are orthogonal and thus the collinearity problem is overcome if we disregard the linear combinations with small variability explanation power.[19] After developing MLPCA, which estimates the PCA subspace in an optimal maximum likelihood sense, when data are affected by measurement errors with a known uncertainty structure,[20] Wentzell et al.[21] applied it in the context of developing a PCR methodology that incorporates measurement uncertainties (MLPCR). As in PCR, MLPCR consists of first estimating a PCA model, now using MLPCA, to calculate the scores through nonorthogonal (maximum likelihood) projections to the estimated MLPCA subspace (instead of the PCA orthogonal projections), and then applying OLS to develop a final predictive model. This technique makes use of the available uncertainty information in the former phases (estimation of a MLPCA model and calculation of its scores), but not during the stage at which OLS is applied. Therefore, Martínez et al.[10] proposed a modification to the regression phase, to make it consistent with the efforts of integrating uncertainty information carried out in the initial stages, which consists of replacing OLS by MLS (we will call this modification MLPCR1). To implement MLS in the second phase, estimated score uncertainties for the ith observation need to be calculated, being given by the diagonal elements of the following matrix[10]

$$Z_i = \{P^T[\mathrm{diag}(uX(i,:))]^{-1}P\}^{-1} \quad (8)$$

where diag is an operator that converts a vector into a diagonal matrix, and P is the matrix of maximum likelihood loads. In our study, we will compare these algorithms based on OLS and MLS (MLPCR and MLPCR1, respectively), with the one obtained when we use the MLMLS algorithm instead of MLS, in the second phase of MLPCR (MLPCR2).

PLS group

PLS[17,18,22-27] is a widely used algorithm in the chemometrics community that also adequately handles noisy data with correlated predictors in the estimation of a linear multivariate model. As in PCR, PLS finds a set of uncorrelated linear combinations of the predictors, belonging to some lower-dimensional subspace in the X-variables space, where y is to be regressed onto. However, in PLS, this subspace is the one that, while still adequately covering the X-variability, provides a good description of the variability exhibited by the Y-variable(s). Here we will make reference to a pair of classes of PLS algorithms, one implemented from raw data and another based on covariance matrices.

Table 3. PLS1 as a Succession of Optimization Subproblems (First Column) and Its Counterparts That Make Use of Information Regarding Measurement Uncertainties

PLS1	*unc-PLS1*	*unc-PLS2*
Step 1. Pre-treatment Center X and y; Scale X and y.	***Step 1. Pre-treatment*** Center X and y ; Scale X and y. Scale X and y uncertainties.	***Step 1. Pre-treatment*** Center X and y; Scale X and y. Scale X and y uncertainties.
Begin For Cycle a=1 : # latent variables	**Begin For Cycle** a=1 : # latent variables	**Begin For Cycle** a=1 : # latent variables
Step 2. Calculate the a^{th} X-weights vector (w) $w = \arg\min_{w} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(X(i,j)-u(i)\times w(j)\right)^2$ $w_{new} \leftarrow w_{old}/\|w_{old}\|$ Note: for a=1, the Y-scores, u, are equal to y.	***Step 2. Calculate the a^{th} X-weights vector (w)*** $w = \arg\min_{w} \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(X(i,j)-u(i)\times w(j)\right)^2}{uX(i,j)^2 + w(j)^2\times uy(i)^2}$ $w_{new} \leftarrow w_{old}/\|w_{old}\|$	***Step 2. Calculate the a^{th} X-weights vector (w)*** $w(j) = \arg\min_{w(j)}\left\{-\frac{1}{2}n\ln(2\pi) - \sum_{i=1}^{n}\ln\left(\sigma_{\varepsilon_{i,j}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\left(X(i,j)-\hat{X}(i,j)\right)^2}{\sigma^2_{\varepsilon_{i,j}}}\right)\right\}$, where, $\sigma^2_{\varepsilon_{i,j}} = (uX(i,j))^2 + (uu(i))^2\, w(j)^2$; $w_{new} \leftarrow w_{old}/\|w_{old}\|$
Step 3. Calculate a^{th} X-scores vector (t) $t = \arg\min_{t} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(X(i,j)-t(i)\times w(j)\right)^2$	***Step 3. Calculate a^{th} X-scores vector (t)*** $t = \arg\min_{t} \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(X(i,j)-t(i)\times w(j)\right)^2}{uX(i,j)^2}$	***Step 3. Calculate a^{th} X-scores vector (t)*** $t = \arg\min_{t} \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(X(i,j)-t(i)\times w(j)\right)^2}{uX(i,j)^2}$
Step 4. Calculate a^{th} X-loadings vector (p) $p = \arg\min_{p} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(X(i,j)-t(i)\times p(j)\right)^2$	***Step 4. Calculate a^{th} X-loadings vector (p)*** $p = \arg\min_{p} \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(X(i,j)-t(i)\times p(j)\right)^2}{uX(i,j)^2 + p(j)^2\times ut(i)^2}$	***Step 4. Calculate a^{th} X-loadings vector (p)*** $p(j) = \arg\min_{p(j)}\left\{-\frac{1}{2}n\ln(2\pi) - \sum_{i=1}^{n}\ln\left(\sigma_{\varepsilon_{i,j}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\left(X(i,j)-\hat{X}(i,j)\right)^2}{\sigma^2_{\varepsilon_{i,j}}}\right)\right\}$, where $\sigma^2_{\varepsilon_{i,j}} = (uX(i,j))^2 + (ut(i))^2\, p(j)^2$
Step 5. Re-scale X-loadings, X-scores and X-weights $p_{new} \leftarrow p_{old}/\|p_{old}\|$; $t_{new} \leftarrow t_{old}\times\|p_{old}\|$; $w_{new} \leftarrow w_{old}\times\|p_{old}\|$	***Step 5. Re-scale X-loadings, X-scores and X-weights*** $p_{new} \leftarrow p_{old}/\|p_{old}\|$; $t_{new} \leftarrow t_{old}\times\|p_{old}\|$; $w_{new} \leftarrow w_{old}\times\|p_{old}\|$ *Step 5.1. Update ut(i), i=1:n.*	***Step 5. Re-scale X-loadings, X-scores and X-weights*** $p_{new} \leftarrow p_{old}/\|p_{old}\|$; $t_{new} \leftarrow t_{old}\times\|p_{old}\|$; $w_{new} \leftarrow w_{old}\times\|p_{old}\|$ *Step 5.1. Update ut(i), i=1:n.*
Step 6. Regression of u on t (b) $b = \arg\min_{b} \sum_{i=1}^{n}\left(u(i)-t(i)\times b\right)^2$.	***Step 6. Regression of u on t (b)*** $b = \arg\min_{b} \sum_{i=1}^{n}\frac{\left(u(i)-b\times t(i)\right)^2}{uu(i)^2 + b^2\times ut(i)^2}$	***Step 6. Regression of u on t (b)*** $b = \arg\min_{b}\left\{-\frac{1}{2}n\ln(2\pi) - \sum_{i=1}^{n}\ln\left(\sigma_{\varepsilon_{i,j}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\left(u(i)-\hat{u}(i)\right)^2}{\sigma^2_{\varepsilon_{i,j}}}\right)\right\}$, where $\sigma^2_{\varepsilon_{i,j}} = (uu(i))^2 + (ut(i))^2\, b^2$
Step 7.Calculation of X and Y residuals $E_a = E_{a-1} - t_a p_a^T \quad (X = E_0)$ $F_a = F_{a-1} - b_a t_a \quad (y = F_0)$ Note: Continue the calculations with E_a playing the role of X and F_a the one of y (u).	***Step 7.Calculation of X and Y residuals*** $E_a = E_{a-1} - t_a p_a^T \quad (X = E_0)$ $F_a = F_{a-1} - b_a t_a \quad (y = F_0)$ *Step 7.1. Up-date* $\{uE(i,j),\ uF(i)\}_{i=1,n;j=1,m}$.	***Step 7.Calculation of X and Y residuals*** $E_a = E_{a-1} - t_a p_a^T \quad (X = E_0)$ $F_a = F_{a-1} - b_a t_a \quad (y = F_0)$ *Step 7.1. Up-date* $\{uE(i,j),\ uF(i)\}_{i=1,n;j=1,m}$.
End For Cycle	**End For Cycle**	**End For Cycle**

Table 4. SIMPLS Algorithm[43]

$S = X^TX$
$s = X^Ty$
for a=1, . . . ,A
r = 1st left singular vector of s
$r = r/(r^TSr)^{1/2}$
$R = [R,r]$
$P = [P,Sr]$
$s = [I - P(P^TP)^{-1}P^T]s$
end
$T = XR$

PLS algorithms implemented directly from raw data

The algorithmic nature of PLS[22,26] can be translated into the solutions of a succession of optimization subproblems,[17,18,23] as presented in the first column of Table 3 for one of its common versions, relative to the case of a single response variable (PLS1). However, if besides having available raw data, $[X \mid y]$, we also know their respective uncertainties, $[uX \mid uy]$, then one way to incorporate this additional information into a PLS algorithm is through an adequate reformulation of the optimization subtasks. Therefore, we have modified the objective functions underlying each optimization subproblem to incorporate measurement uncertainties, but still preserving the successful algorithmic structure of PLS. Such a sequence of optimization subproblems is presented in the second and third columns of Table 3, where MLS and MLMLS replace OLS in several algorithmic stages, giving rise to the uncertainty-based equivalents unc-PLS1 and unc-PLS2, respectively.

PLS algorithms implemented from covariance matrices

There are several alternative ways to develop a PLS model, most of them leading to very similar or even exactly the same results. In fact, Helland[25] has shown the equivalence between two of such algorithms (one based on orthogonal scores and another using orthogonal loadings instead), both of them based on available raw data matrices for the predictors and response variables. Another class of PLS methods that encompasses the so-called SIMPLS, developed by Sijmen de Jong (see Table 4), or the approach presented by Kaspar and Ray,[28] built on previous work from Höskuldsson,[29] consists of algorithms entirely based on data covariance or cross-product matrices. For the single response case, a SIMPLS solution provides exactly the same results as Svant Wold's orthogonalized PLS algorithm, leading to only minor differences when several outputs are considered. Matrices S and s in Table 4 do play a central role in PLS. Theoretical analysis of this algorithm[25,30] leads to the conclusion that the calculated vector of coefficients, when a latent variables are considered, $\hat{\beta}^a_{PLS}$, is given by

$$\hat{\beta}^a_{PLS} = V_a(V_a^TSV_a)^{-1}V_a^Ts \tag{9}$$

where $V_a = [v_1, v_2, \ldots, v_a]$ is any $(m \times a)$ matrix whose columns span the following Krylov subspace, $\Re^a(s; S)$, that is, the subspace generated by the first a columns of the Krylov sequence, $\{s, Ss, \ldots, S^{a-1}s\}$. Thus, matrices S and s define the structure of the relevant Krylov subspace where the PLS solution will lie. In fact, the columns of the PLS weighting matrix W, which define the subspace of the full predictor space with maximal covariance with the response, do form an orthogonal base of $\Re^a(s; S)$. The relevancy of S and s for PLS provided the motivation to direct some efforts toward the incorporation of uncertainty information in the computation of better *estimates* for both of these matrices. The reason that we have not called them estimates so far is explained by the lack of a consistent statistical population model underlying PLS.[24,31,32] However, when we now say that our goal is to calculate "better" covariance matrices, this implies that some goodness criteria must be assumed. Therefore, to give a step forward toward the integration of measurement uncertainties in our analysis, one should postulate a statistical model to provide an estimation setting for the covariance matrices S and s. For the sake of the present work, we consider the following latent variable multivariate linear relationship for $Z = [x^T \mid y]^T$, which has the ability to incorporate heteroscedastic measurement errors with known uncertainties (these uncertainties are considered by now to be independent of the true levels for the noiseless measurands)

$$Z(k) = \mu_Z + A \cdot l(k) + \varepsilon_m(k) \tag{10}$$

where Z is the $(m + 1) \times 1$ vector of measurements, μ_Z is the $(m + 1) \times 1$ mean vector of x, A is the $(m + 1) \times a$ matrix of model coefficients, l is the $a \times 1$ vector of latent variables, and $\bar{\varepsilon}_m$ is the $(m + 1) \times 1$ vector of measurement noise. This model is still incomplete because we need to provide it with the probability density functions assumed for each random component

$$l(k) \sim \text{iid } MN_a(0, \Delta_l),\ \varepsilon_m(k) \sim \text{id } MN_{m+1}[0, \Delta_m(k)]$$

$$l(k) \quad \text{and} \quad \varepsilon_{m(j)} \quad \text{are independent } \forall k, j \tag{11}$$

where MN stands for multivariate normal distribution, Δ_i is the covariance matrix of the latent variables, $\Delta_m(k)$ is the covariance matrix of the measurement noise at time k, given by $\Delta_m(k) = \text{diag}[\sigma^2_m(k)]$. Thus, for estimating the covariance matrix, we assume a multivariate behavior for Z that can be adequately described by propagation of the underlying variation of p latent variables, plus added noise in the full variable space. This model and the calculation details associated with the estimation of the unknown parameters are fully described elsewhere.[33] It can be shown that the probability density function of Z, under the conditions stated above, is a multivariate normal distribution with the following form

$$Z(k) \sim \text{id } MN_{m+1}[\mu_Z, \Sigma_Z(k)] \tag{12}$$

where

$$\Sigma_Z(k) = \Sigma_l + \Delta_m(k) \qquad \Sigma_l = A\Delta_lA^T \tag{13}$$

With the raw measurements (Z) and the associated uncertainties [from which we can calculate $\Delta_m(k)$], it is possible to estimate μ_Z and Σ_l by maximizing the likelihood function. Matrix $\Sigma_Z(k) = \Sigma_l + \Delta_m(k)$ is the estimate of the covariance matrix for noisy measurements at time step (k), but because PLS is based on S and s, it requires single estimates for the population parameters (and not one per time step k). Thus, we maintain the estimate of the covariance of noiseless data, $\hat{\Sigma}_l$,

but average out the heteroscedastic square uncertainties, to come up with a single term, $\hat{\Delta}_m$, leading to

$$\hat{\Sigma}_Z \cong \hat{\Sigma}_l + \overline{\hat{\Delta}}_m \tag{14}$$

With the estimate of Σ_Z, we can finally calculate the estimates for S and s: $S = \hat{\Sigma}_Z(1:m, 1:m)$, $s = \hat{\Sigma}_Z(1:m, m+1)$. The algorithm that consists of implementing the SIMPLS algorithm with these matrices as inputs will be referred to here as unc-PLS3. In the present context, we use the full measurement space to estimate Σ_Z ($a = m$) because we want the relevant subspace for prediction to be defined by the PLS algorithm itself, and not by a previous estimation step. In the *prediction phase*, when new values for the predictors become available along with their measurement uncertainties, and the goal is to predict what the value of the response variable would be, we add an additional calculation step *before* applying the unc-PLS3 regression vector (calculated in the *estimation phase*). This step consists of projecting the new multivariate observation in the full X-space into the subspace that is relevant for predictions (that is, the one spanned by the columns of the weighting matrix, W in PLS *or* R in SIMPLS). The availability of the associated uncertainties leads to a generally nonorthogonal projection methodology that consists of estimating the projected points using a maximum likelihood approach, just as the one adopted in MLPCA.[21] In the present study, we also tested an algorithm that implements the same nonorthogonal projection operation, but using the weighting matrix provided by PLS (a hybrid version of the classic PLS because it contains a projection step that incorporates measurement uncertainty), herein referred to as unc-PLS4. For the sake of completeness, we also introduced another methodology, based on the same weighting matrix as unc-PLS3, but that bypasses the nonorthogonal projection step, designated as unc-PLS5.

Measurement Uncertainties in Process Optimization

In the previous section we have addressed the explicit incorporation of measurement uncertainty in statistical model development. We now move to a different working scenario, where an appropriate model is already available and our goal is to use it for process optimization, but also taking into account information regarding measurement and actuation uncertainties. In particular, we address the problem where one wants to optimize an objective function (such as maximizing some profit metric or minimizing a cost function), for a given measurement of the vector of load variables (L), by manipulating another set of variables (M). However, because of the presence of uncertainties, the following issues do arise:

- *Measured* quantities (that is, the loads $\tilde{L}$ and the outputs $\tilde{Y}$) are affected by measurement noise, with statistical characteristics defined by their associated uncertainty

$$\tilde{L} = L + \varepsilon_L \qquad \tilde{Y} = Y + \varepsilon_Y \tag{15}$$

with quantities marked with a tilde accent (~) being the values actually available, whereas L and Y are the corresponding true, but unknown, values for these quantities (Figure 1).

- Similarly, the set-point that we specify for the manipulated variables ($\tilde{Z}$) does not correspond to the exact true value of the manipulation action over the process. In fact, because of *actuation noise*, there is also here another uncertainty source to be taken into account.

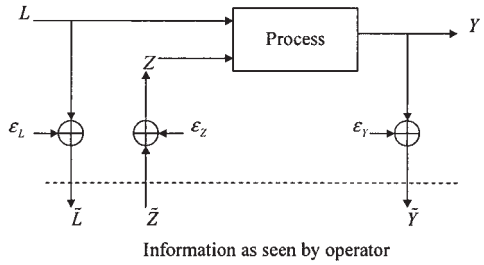

Figure 1. Schematic representation of measured quantities [as seen by an external operator and marked with a tilde (~)] and the quantities that are actually involved in the underlying process.

Considering that we want to drive the process in such a way as to minimize some relevant cost function, $\phi(\cdot)$, we propose the following formulation that incorporates measurement and actuation uncertainties, in the calculation of the adequate values for the manipulated variables to be specified externally, when a given measurement for the load is acquired ($\tilde{L}$). As often happens in the formulation of optimization problems under uncertainty, the objective function constitutes an expected value for the performance metric, taken over the space of uncertain parameters:

Formulation I

$$\begin{aligned} &\min_{\tilde{Z}} E_\Theta\{\phi(L, Z, \tilde{Y})\} \\ &\text{s.t. } g(Y, L, Z) = 0 \\ &\quad L = \tilde{L} - \varepsilon_L \\ &\quad Z = \tilde{Z} + \varepsilon_Z \\ &\quad \tilde{Y} = Y + \varepsilon_Y \end{aligned} \tag{16}$$

where $E_\Theta\{\cdot\}$ is the expectation operator,

$$E_\Theta\{\phi\} = \int_\Theta \phi(\theta) j(\theta) d\theta \tag{17}$$

$\theta = [\varepsilon_L^T, \varepsilon_Z^T, \varepsilon_Y^T]^T$ and $j(\theta)$ provide the joint probability density function for the uncertain quantities θ. The available model is represented by $g(Y, L, Z) = 0$, and we will assume here that the uncertainty associated with its parameters is negligible (if not, such uncertainties can also be incorporated into our problem formulation[34]).

In Formulation I, we assume that the relevant quantities for evaluation of the performance metric are the values of L and Z that really affect the process, as well as the *measured* value of the output. We point out that these assumptions do not necessarily hold in every situation. For instance, sometimes the performance metric should be calculated with the "true" value of the output, Y, instead of $\tilde{Y}$] (Formulation II, see below), as is the case when output measurements become available with much less uncertainty in a subsequent stage (such as from

off-line laboratory tests). Other times, only measured values should be used because no better measurements or reconciliation procedures can be adopted. The correct formulation is therefore case dependent, and should be tailored to each particular situation.

Formulation II

$$\begin{aligned} &\min_{\tilde{Z}} E_{\Theta}\{\phi(L, Z, Y)\} \\ &\text{s.t. } g(Y, L, Z) = 0 \\ &\quad L = \tilde{L} - \varepsilon_L \\ &\quad Z = \tilde{Z} + \varepsilon_Z \end{aligned} \tag{18}$$

In our case studies section we will also present the results obtained for the situation where uncertainties are not at all taken into account, and thus where the manipulated variable values are found by solving the following problem

Formulation III

$$\begin{aligned} &\min_{\tilde{Z}} \phi(\tilde{L}, \tilde{Z}, \tilde{Y}) \\ &\text{s.t. } g(\tilde{Y}, \tilde{L}, \tilde{Z}) = 0 \end{aligned} \tag{19}$$

Case Studies

In this section we present the results reached from comparative analysis encompassing all the methods mentioned above (PLS, unc-PLS1, unc-PLS2, unc-PLS3, unc-PLS4, unc-PLS5, RR, rMLS, rMLMLS, PCR, MLPCR, MLPCR1, MLPCR2, OLS, MLS, and MLMLS), and illustrate the implementation of the approach treated in the third section under a realistic simulation scenario, using a model estimated from a real paper pulp pilot digester.

Case studies 1 and 2 provide different contexts to set a ground for the comparison study among the multivariate linear regression methods. In both of them, a latent variable model structure is adopted to generate simulated data, given that this kind of model structure is quite representative of data collected from many real industrial processes because the number of inner sources of variability that drives process behavior is usually of a much smaller dimensionality than the number of measured variables.[35,36] The latent variable model used has the following form

$$\begin{aligned} X &= \mathbf{1}_n \cdot \mu_X^T + TP + E \\ Y &= \mathbf{1}_n \cdot \mu_Y^T + TQ + F \end{aligned} \tag{20}$$

where μ_X and μ_Y are the $m \times 1$ and $k \times 1$ vectors with the column averages of X and Y; $\mathbf{1}_n$ is an $n \times 1$ vector of ones; X is the $n \times m$ matrix of input data; Y is the $n \times k$ matrix of output data; T is the $n \times a$ matrix of latent variables that constitute the inner variability source, structuring both the input and output data matrices; E and F are $n \times m$ and $n \times k$ matrices of random errors; and P and Q are $a \times m$ and $a \times k$ matrices of coefficients.

The model used in our simulations consists of five latent variables ($a = 5$) that follow a multivariate normal distribution with zero means and a diagonal covariance (I_a, that is, the identity matrix of dimension a). The dimension of the input space is set equal to 10 and that of the output space equal to 1 ($m = 10$, $k = 1$). Rows of the P matrix form an a-orthonormal set of vectors with dimension m. The same applies to matrix Q, which consists of an a-orthonormal set of vectors with dimension k.

Each element of matrices E and F of random errors is drawn from a normal distribution with zero mean and standard deviation given by the uncertainty level associated with that specific variable (column of X or Y) for a particular observation (row). These uncertainties were allowed to vary, and this variation is characterized by the heterogeneity level (HLEV), which measures the degree of variation or heterogeneity of uncertainties from observation to observation: HLEV = 1 means a low variation of the noise uncertainty or standard deviation from observation to observation, whereas HLEV = 2 means a highly heteroscedastic behavior for the noise uncertainties. More specifically, for variable X_i the uncertainties along the observation index are randomly generated from a uniform distribution centered at $\bar{u}(X_i)$ (the average uncertainty for a given variable), with range given by $R(HLEV) = K_2(HLEV) \times \bar{u}(X_i)$, where $K_2 = 0.01$ (if HLEV = 1; low heterogeneity level) or $K_2 = 1$ (if HLEV = 2; high heterogeneity level), that is,

$$u[X_i(k)] \sim U\left[\bar{u}(X_i) - \frac{R(HLEV)}{2}, \bar{u}(X_i) + \frac{R(HLEV)}{2}\right]$$

In the present study, $\bar{u}(X_i)$ was kept constant at 0.5 times the theoretical standard deviation calculated for each noiseless variable.

Case study 1: complete heteroscedastic noise

With the goal of evaluating overall performance of the methods under different uncertainty structures for the measurements errors, the following sequence of steps was adopted:

(1) We set the tuning parameters for each method and for each set of conditions (number of latent dimensions for PLS and PCR methods, and ridge parameter for RR methods). Regarding PLS and PCR methods, we did set $a = 5$. As for ridge methods, we selected our ridge parameter using cross-validation and the generation of a logarithmic grid in the range of plausible values (the criterion used in cross-validation is RMSEPW). This procedure is repeated 10 times, and the median of the best values is chosen as the tuning parameter to be used in our simulations. Variables are "auto-scaled" in all methods, except for OLS, MLS, and MLMLS.

(2) For each scenario of HLEV (1 or 2), two noiseless data sets are generated according to the latent variable model presented above: a training or reference noiseless data set and a test noiseless data set, both with 100 multivariate observations. Furthermore, a random sequence of uncertainties (noise standard deviations) for all the observations belonging to each variable is generated according to HLEV.

(3) Zero-mean Gaussian noise, with standard deviation given by the uncertainties calculated in (2), is generated and added to the noiseless training and testing data sets, after which a model is estimated according to each linear regression method (using the training data set) and its prediction perfor-

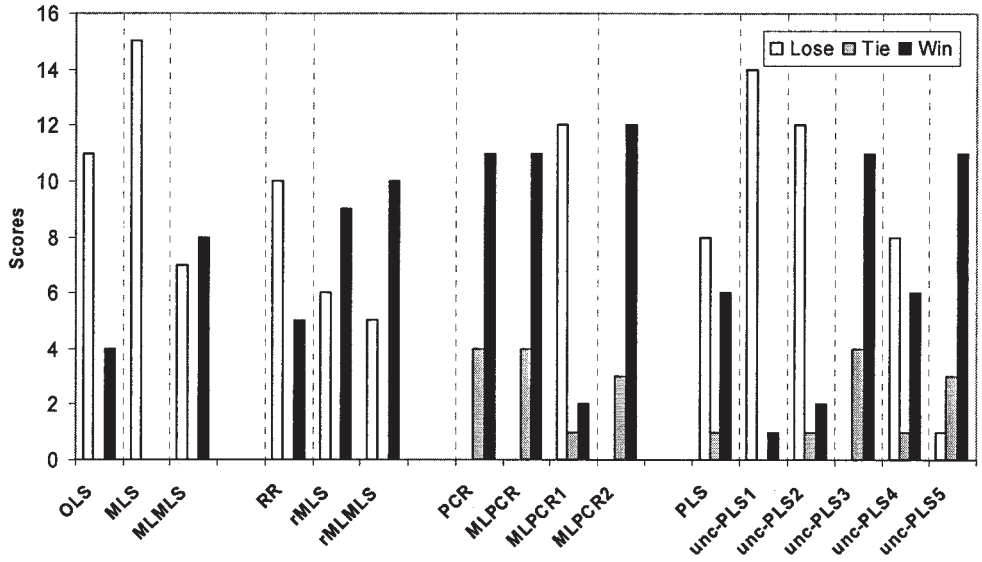

Figure 2. Results for number of losses, ties, and wins for each method, under the simulation scenario with heterogeneity level (HLEV) = 1 [using root mean square error of prediction (RMSEP)].

mance evaluated (using the test data set). This process of noise addition, followed by parameter estimation and prediction, is repeated 100 times, and the corresponding performance metrics saved for future analysis.

Performance metrics used for prediction assessment are the square root of the weighted mean square error of prediction in the test set (RMSEPW), where the weights are the result of combining the predictor and response uncertainties, and the more familiar root mean square error of prediction (RMSEP)

$$RMSEPW(i) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\frac{[y(k)-\hat{y}(k)]^2}{uy(k)^2 + [uX(k,:)^{*2}]^T B^{*2}}} \qquad i = 1100 \tag{21}$$

$$RMSEP(i) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}[y(k)-\hat{y}(k)]^2} \qquad i = 1100 \tag{22}$$

where n is the number of observations in the test set.

At the end of the simulations, we do have 100 values for the above metrics available for comparing the performances achieved by the different methods, under a given noise structure scenario. To take into account both the individual variability of the performance metrics for the different methods, as well as their mutual correlations, we based the comparison strategy in paired t-tests among all the different combinations of methods. Therefore, for each simulation scenario, paired t-tests were used to determine whether method A is better than method B (a Win for method A), performs worse (a Loss), or if there is no statistical significant difference between both of methods A and B (a Tie), for a given significance level (we used $\alpha = 0.01$). For the sake of simplicity, we will only present here the number of wins, losses, and ties that each method obtained for each simulation scenario.

Figure 2 presents the comparison results for the scenario HLEV = 1, using RMSEP as performance metric (because the trends for RMSEP and RMSEPW do not differ significantly, only those for the more familiar RMSEP are presented).

Examining first the performance of the methods belonging to the same group, we can see the following for this simulation scenario:

- OLS Group. MLS performs worse than OLS and MLMLS shows the best performance among the three methods. In general terms, comparing all the methods where MLS and MLMLS have similar roles (such as unc-PLS1/unc-PLS2, rMLS/rMLMLS, MLPCR1/MLPCR2), the second version never resulted in worse results and, as a matter of fact, almost always significantly improved them.
- RR Group. Both rMLS and rMLMLS conducted to improved results with respect to those obtained by RR.
- PCR Group. MLPCR does not improve over PCR predictive results, but MLPCR2 leads to an improvement.
- PLS Group. Methods unc-PLS3 and unc-PLS5, both using uncertainty-based estimation of the relevant covariance matrices for PLS, present the best performance. Their similar performance results can be explained by the fact that, under mild homoscedastic situations and if the variables present approximately equal uncertainties associated with them, the orthogonal and nonorthogonal projections almost coincide. The same applies for the comparison of PLS and unc-PLS4, both using PLS weighting vectors but different projection strategies. Comparing the results obtained for all the methods against each other, we can see that MLPCR2 is the one that presented the best overall performance, followed by PCR, MLPCR, unc-PLS3, and unc-PLS5.

Figure 3 summarizes the results obtained for condition HLEV = 2. A comparison of performances regarding methods within the PLS group shows that those methods that estimate the covariance matrices using uncertainty information (unc-PLS3, unc-PLS5) present better performance then their counterparts that use the same projection strategies (unc-PLS4, PLS, respectively). However, looking now to the methods that differ

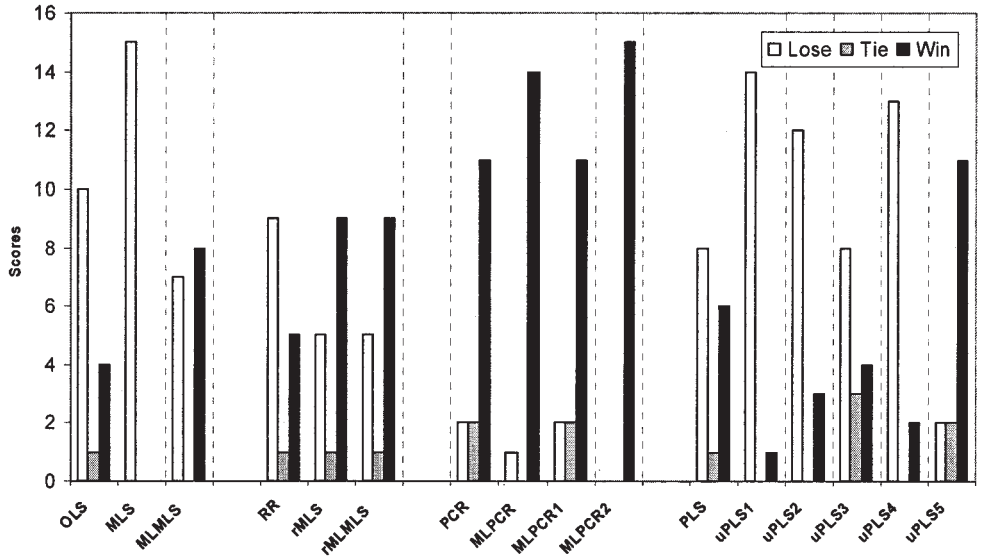

Figure 3. Results for number of losses, ties, and wins for each method, under the simulation scenario with HLEV = 2 (using RMSEP).

only on the projection methodology, we can see that those that are based on orthogonal projections achieve better results that those based on nonorthogonal maximum likelihood projections. This result is quite interesting and will be further discussed below. In the PCR group we can see that all the methods perform quite well. As for the remaining groups of methods, the trends mentioned for HLEV = 1 remain roughly valid. MLPCR2 continues to be the method with the best overall performance, followed by MLPCR and a group of methods that include MLPCR1, PCR, and unc-PLS5.

Case study 2: handling missing data

In this second case study, we analyze the prediction performance of the several methods when missing data are present (both in model estimation and in prediction), and a very simple strategy for handling missing data is adopted: mean substitution. For uncertainty-based methods, one also has to specify the associated uncertainty, and the values we have considered here are the standard deviations of the respective variables during normal operation. Other more sophisticated methodologies for missing data imputation during model estimation are also available for regression methods (especially PLS and PCR[37]), as well as methods for handling missing data once we have already available an estimated model.[38] Analogous approaches can also be developed for the uncertainty-based techniques that require only the estimated value and the respective uncertainty to fill existing blanks. However, the aim of this study is to assess the extent to which one can easily handle missing data in model estimation and prediction (that is, with minimum assumptions regarding missing values and the least modification over standard procedures), taking advantage of the possibility of using uncertainty information. That being the case, we decided to keep the same replacement strategy among all methods, so that the real advantage of handling such an additional piece of information, provided by measurement uncertainties, can be easily evaluated and compared with the current alternatives.

Because our focus here is related with the evaluation of the methods regarding prediction when missing data is present, we adopted a simulation structure which is now different from that of case study 1. For *each simulation* the following steps are repeated and the corresponding results saved:

(1) Generate a new latent variable model (matrices Q and P) and noiseless data to be used for model estimation and prediction assessment. Also generate measurement uncertainties to be associated with each nonmissing value, according to the value of *HLEV* used in each simulation study.

(2) Generate a new "missing data mask" that removes (on average) a chosen percentage of the data matrix $[X \mid Y]$. We used a target percentage of 20%, both for the reference and test data sets.

(3) Generate and add noise to the noiseless data that were not removed, according to the measurement uncertainties generated in (1).

(4) Replace missing data with column means for the data set used to estimate the model, and calculate the associated uncertainties using the columns standard deviations, for the same data set.

(5) Estimate models using the data set constructed in (4).

(6) For the test data set, do the same operation as in (4). (using the same values for the input values and uncertainties) and calculate the predicted value for the output variable. Calculate overall performance metrics (RMSEPW and RMSEP).

The results obtained with HLEV = 1 are presented in Figure 4, where we can see that within the PLS group methods unc-PLS5 and unc-PLS3 lead to improved predictive performances, but now with unc-PLS3 presenting better results that unc-PLS5, that is, the nonorthogonal projection seems to bring some added value when missing data are present, under homoscedastic scenarios. In the PCR group, all MLPCR methods outperform the conventional PCR. As for the other groups, results obtained follow the same trends verified when no missing data were present. In global terms, MLPCR2 presents the

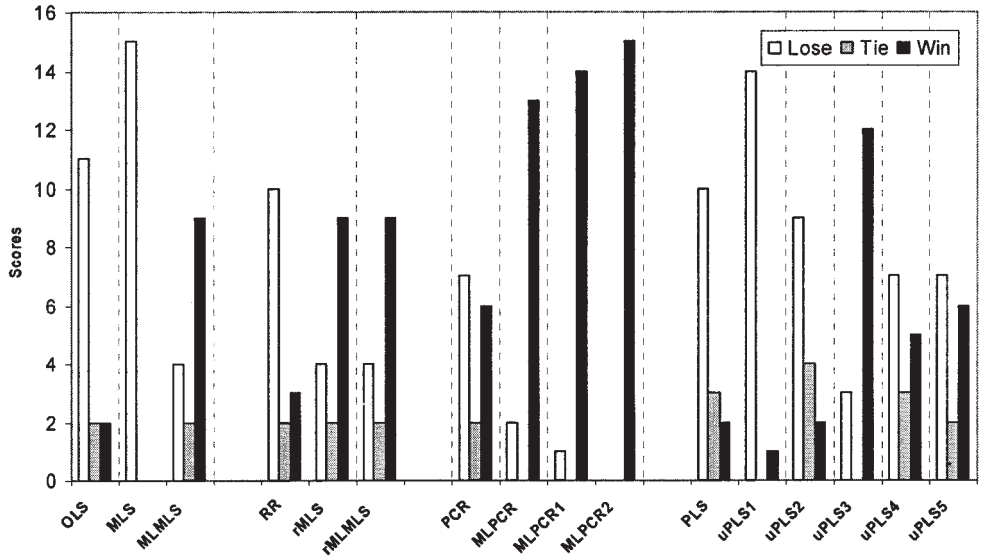

Figure 4. Results for number of losses, ties, and wins for each method, under the simulation scenario with HLEV = 1 and 20% of missing data (using RMSEP).

best overall performance, followed by MLPCR1, MLPCR, and unc-PLS3.

By analyzing the results for HLEV = 2 (Figure 5), we can also see that unc-PLS3 and unc-PLS5 still show the best predictive performance within the PLS group, but now with unc-PLS3 presenting lower scores relatively to the previous scenario (HLEV = 1), a result that is consistent with what was verified in case study 1. In the global comparison, after MLPCR2 we can find MLPCR1 and MLPCR. Therefore, under the conditions adopted for this simulation study, we can conclude that MLPCR methods tend to have the best overall performance in the presence of missing data.

We point out that when adopting a methodology that integrates data uncertainty, one follows the same calculation procedure adopted for the situation where no data are missing, simply replacing the missing elements with rough estimates that will be properly weighted by the algorithms, according to their associated uncertainties. However, if we do have available better estimates, such as those arising from more sophisticated imputation techniques, one can also integrate them as well, without any further changes.

Case study 3: process optimization under data uncertainty

This case study illustrates the integration of measurement uncertainties in process optimization decision making. The problem we address herein consists of calculating the values

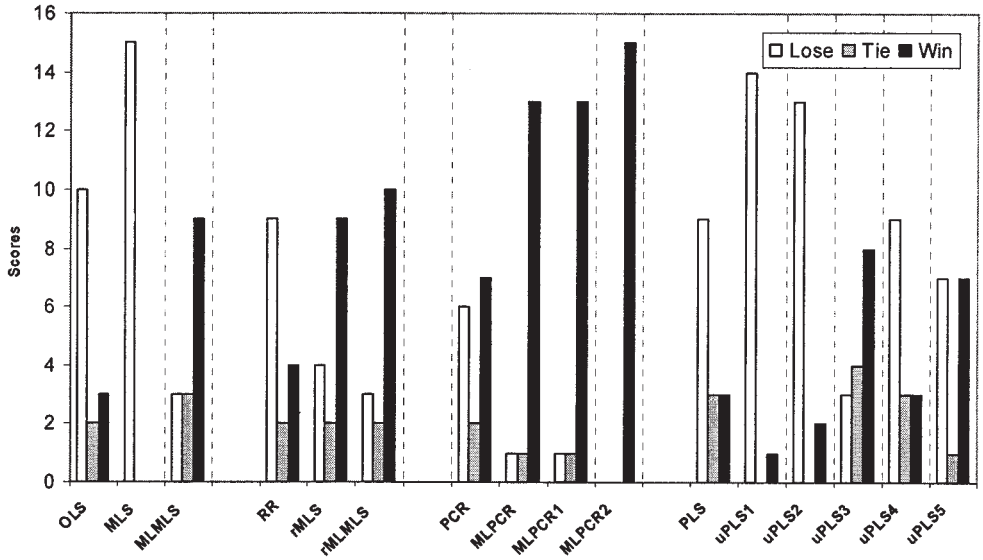

Figure 5. Results for number of losses, ties, and wins for each method, under the simulation scenario with HLEV = 2 and 20% of missing data (using RMSEP).

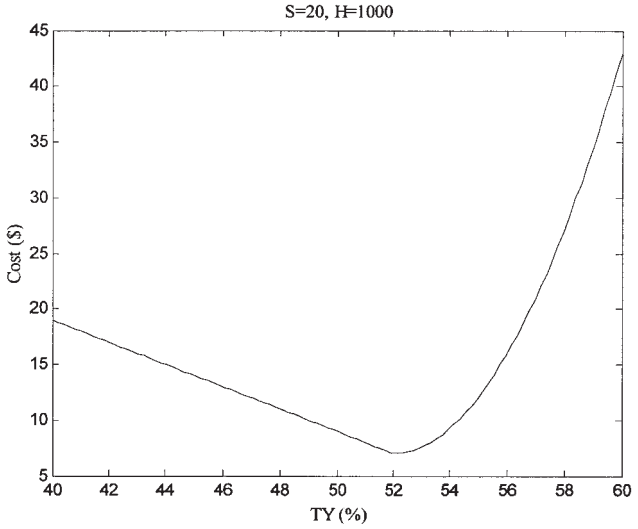

Figure 6. Cost function for deviations of total yield (TY) from its target value (52%), for $S = 20$ and $H = 1000$.

for the manipulated variables to be specified ($\tilde{Z}$) to minimize a cost function, when measurements for the loads become available ($\tilde{L}$). This particular case study is based on the following model, developed for a batch paper pulp pilot digester[39]

$$TY = 55.2 - 0.39 \times EA + 324/(EA \times \log_{10}S) - 92.8 \times \log_{10}(H)/(EA \times \log_{10}S) \quad (23)$$

This model relates pulp total yield (TY) with effective alkali (EA, a measure of the joint concentration of Na_2OH and Na_2S, the active elements in the cooking liquor), sulfidity (S, the percentage of Na_2S in the cooking liquor), and H factor (H, a function of the temperature profile across the batch).

We consider the situation where a cost function (L) penalizes deviations from a target value for TY (52%): the penalty for lower values is attributed to fiber losses, and that for higher values to deterioration in other pulp properties. Our cost function also considers the cost of S and H (proportional to their respective magnitudes). As an example, Figure 6 illustrates the shape of the assumed cost function for $S = 20$ and $H = 1000$

$$L = \begin{cases} 100\left(\dfrac{TY_{sp}}{100} - \dfrac{TY}{100}\right) + \dfrac{S}{4} + \dfrac{H}{500} & \Leftarrow \quad TY \le TY_{sp} \\ 75^2\left(\dfrac{TY_{sp}}{100} - \dfrac{TY}{100}\right)^2 + \dfrac{S}{4} + \dfrac{H}{500} & \Leftarrow \quad TY > TY_{sp} \end{cases} \quad (24)$$

Table 6. Solutions Obtained under Formulations I, II, and III, and Their Associated Average Costs

Solutions	Average Cost ($)	
	Formulation I	Formulation II
I $\tilde{S}=$ 7.16 $\tilde{H}=$1602.0	10.80	5.93
II $\tilde{S}=$ 7.83 $\tilde{H}=$1184.2	11.16	5.40
III $\tilde{S}=$ 5.38 $\tilde{H}=$1274.6	25.46	8.17

In this example, EA is assumed to be a load variable, and thus our optimization goal consists of calculating the S and H values that minimize expected cost in the presence of uncertainties for both measurements and process actuations. Formulations I, II, and III hold for this example, with $L = EA$, $Z = [S \quad H]$, and $Y = TY$ (Table 5).

We further assumed that the vector of uncertain quantities, $\theta = [\varepsilon_{EA}, \varepsilon_S, \varepsilon_H, \varepsilon_{TY}]^T$, follows a multivariate normal distribution with zero mean and diagonal covariance given by

$$\Sigma_\Theta = \text{diag}([2^2 \quad 2^2 \quad 50^2 \quad 4^2]) \quad (25)$$

where diag stands for the operator that converts a vector into a diagonal matrix with its elements along the main diagonal.

To illustrate the implementation of the formulations above referred, let us consider that the observed value for EA is 15 ($\widetilde{EA}$). Table 6 summarizes the results obtained for the manipulated variables ($\tilde{S}$ and $\tilde{H}$) and the average cost obtained with the objective function assumed under formulations I and II, with a third degree specialized cubature being used for estimation of expected values.[40]

From Table 6 we can see that under the simulation conditions considered here, and assuming that the relevant objective function is the one associated with formulation I, the optimal solution obtained when one disregards measurement and actuation uncertainties (formulation III) corresponds to an average cost increased by 136%. If the relevant objective function were the one corresponding to problem formulation II, the average cost increase would be 51%. It should also be noticed that the location of the optimal solution in the ($\tilde{S}$, $\tilde{H}$) decision space, found if one ignores uncertainties, is quite distant from the true one.

The cost associated with the nonconsideration of these types of uncertainties decreases when their magnitude becomes smaller. Figure 7 presents the results obtained for three alternative problem formulations, when the covariance matrix for uncertain quantities is multiplied by a monotonically decreas-

Table 5. Optimization Formulations I, II, and III as Applied to Case Study 3

Formulation I	Formulation II	Formulation III
$\min_{S,H} E_\Theta\{\phi(EA, S, H, \widetilde{TY})\}$	$\min_{S,H} E_\Theta\{\phi(EA, S, H, TY)\}$	$\min_{S,H} \phi(\widetilde{EA}, \tilde{S}, \tilde{H}, \widetilde{TY})$
s.t. $g(TY, EA, S, H) = 0$	s.t. $g(TY, EA, S, H) = 0$	s.t. $g(\widetilde{TY}, \widetilde{EA}, \tilde{S}, \tilde{H}) = 0$
$EA = \widetilde{EA} - \varepsilon_{EA}$	$EA = \widetilde{EA} - \varepsilon_{EA}$	
$S = \tilde{S} + \varepsilon_S$	$S = \tilde{S} + \varepsilon_S$	
$H = \tilde{H} + \varepsilon_H$	$H = \tilde{H} + \varepsilon_H$	
$\widetilde{TY} = TY + \varepsilon_{TY}$		

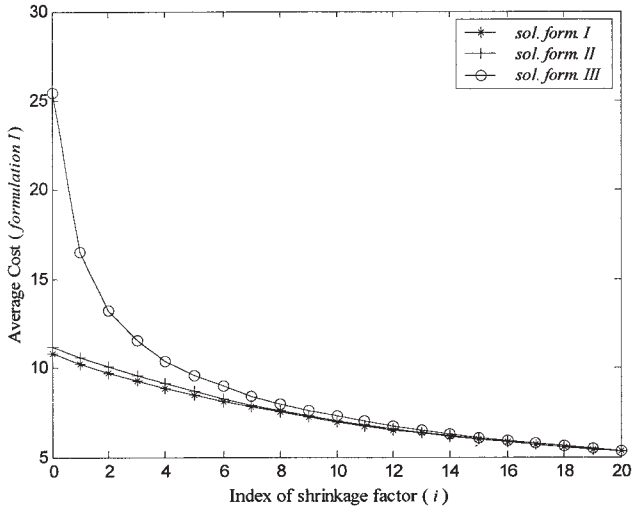

Figure 7. Behavior of average cost (formulation I), corresponding to solutions for the three alternative problem formulations, using $\Sigma_\Theta \cdot 0.9^i$.

ing shrinkage factor, 0.9^i. As expected, the differences arising from the solutions associated with such three optimization formulations tend to vanish when measurement and actuation uncertainties decrease. Furthermore, the average cost also decreases because of the improved quality of information obtained from measurement devices and the better performance of final control elements, as one moves across the several simulation scenarios considered here.

Discussion

Results presented in the previous section highlight not only the potential of using all the information that is available (data and associated uncertainties), but also the difficulty that such a task may encompass, with respect to model estimation. In fact, we have come across with some unexpected results and relevant issues have been identified and merit being discussed here.

First of all, we stress the fact that, even though simulation results are strictly valid within the conditions established, they can provide useful guidelines for real processes that present structural similarities with them. The fact that classical methods do not make explicit use of uncertainty information may not be very relevant if it represents just a small part of the global variability exhibited by variables. Therefore, uncertainty-based methods presented here are expected to bring potentially more added value only under contexts where uncertainty is quite high (noisy environments) or experiments have large variations. In other words, these methods should complement their classical counterparts, depending on the noise characteristics that prevail in measured data.

Still regarding model estimation, we have found some convergence problems in MLMLS, something that is not unusual in approaches based on numerical optimization of a nonlinear objective function. However, problems in MLPCR2 arising from the nonconvergence of MLMLS are usually rare. From the experience that we have gathered so far, no limitations were found regarding the implementation of MLPCR2 in the analysis of real industrial data. The poor performance of MLS under the scenarios considered here, where predictors are strongly correlated, may indicate that the inversion operation undertaken at each iteration is interfering with its performance (the matrix to be inverted in this method becomes quite ill-conditioned under collinear situations of the predictors). Results obtained for the ridge regularization of MLS (rMLS) show an effective stabilization of this operation. As for PLS methods, the extensive solution of small optimization problems can make unc-PLS1 and unc-PLS2 more prone to numerical convergence problems than the original PLS method, something that does not occur with the remaining uncertainty-based PLS methods (unc-PLS3, unc-PLS4, and unc-PLS5), given that they are based on the estimation of covariance matrices and projection operations. Quite interesting is the fact that, when comparing under heteroscedastic situations (Figure 3) PLS methods that adopt the same estimation procedure for the covariance matrices but differ in the projection phase (as happens with pairs PLS/unc-PLS4, unc-PLS3/unc-PLS5), one can see that the use of uncertainty-based maximum-likelihood non-orthogonal projections seems to be detrimental for prediction with respect to orthogonal projections. In fact, a separate simulation study showed evidence toward a reduced variance of the orthogonal projection scores, when compared to the one exhibited by maximum likelihood projection scores. Apparently, for heteroscedastic scenarios, oscillations in the non-orthogonal projection line may also bring some added variability to the scores, other than the one strictly arising from variability attributed to noise sources. This increased dispersion in the reduced space of the scores, usually the one relevant for prediction purposes, can increase prediction uncertainty arising from poorly estimated models, something that is in line with the results presented in Figure 3. Finally, there are also some approximations considered in the methods that may interfere with their predictive performance and should be considered in future developments. That is, methods unc-PLS1 and unc-PLS2 neglect uncertainties in the load vectors and MLPCR1/MLPCR2 do assume the score uncertainties to be independent.

We emphasize that, although we have focused here on steady-state applications, our approaches can also be used under the context of dynamic models, that is, through the consideration of lagged variables[41-44] (the PLS methods based on the uncertainty-based estimation of covariance matrices, however, do need some modifications to cope with the noise correlations appearing with the use of lagged variables). For such situations, one may also consider uncertainty descriptions connected with robust control methodologies, such as *H*-infinity approaches.

Conclusions

In this paper we address the importance of specifying measurement uncertainties and how this information can be used in two distinct tasks: model estimation and process optimization. With respect to model estimation, under the conditions studied method MLPCR2 presented the best overall predictive performance. In general, those methods based on MLMLS present improvements over their counterparts based on MLS. We have also illustrated the potential advantage of using measurement and actuation uncertainties in process optimization problem formulations and solutions. Our study points out the relevance of not neglecting measurement/manipulation uncertainties

when addressing both on-line and off-line process optimization.

Future work will address the application of uncertainty-based methods in real industrial contexts, using the guidelines extracted from the results achieved in our comparative study presented herein, regarding the most adequate methods to be adopted for a certain noise/data structure scenario.

Acknowledgments

The authors gratefully acknowledge FCT (Fundação para a Ciência e Tecnologia, Portugal) for financial support through research project POCTI/EQU/47638/2002.

Literature Cited

1. MacGregor JF, Kourti T. Statistical process control of multivariate processes. *Control Eng Pract.* 1995;3:403-414.
2. Wise BW, Gallagher NB. The process chemometrics approach to process monitoring and fault detection. *J Process Control.* 1996;6: 329-348.
3. Venkatasubramanian V, Yin RRK, Kavuri SN. A review of process fault detection and diagnosis. Parts I–III. *Comput Chem Eng.* 2003; 27:Part I: 293-311, Part II: 313-326, Part III: 327-346.
4. Wang XZ. *Data Mining and Knowledge Discovery for Process Monitoring and Control.* London: Springer-Verlag; 1999.
5. ISO. *Guide to the Expression of Uncertainty.* Geneva, Switzerland: International Organization for Standardization; 1993.
6. Albuquerque JS, Biegler LT. Data reconciliation and gross-error detection for dynamic systems. *AIChE J.* 1996;42:2841–2856.
7. Crowe CM. Data reconciliation—Progress and challenges. *J Process Control.* 1996;6:89-98.
8. Romagnoli JA, Sanchez MC. *Data Processing and Reconciliation for Chemical Process Operation.* Vol. 2. San Diego, CA: Academic Press; 2000.
9. Reis MS, Saraiva PM. A comparative study of linear regression methods in noisy environments. *J Chemomet.* 2004;18(12):526.
10. Martínez À, Riu J, Rius FX. Application of the multivariate least squares regression method to PCR and maximum likelihood PCR techniques. *J Chemomet.* 2002;16:189-197.
11. Río FJ, Rio J, Rius FX. Prediction intervals in linear regression taking into account errors in both axes. *J Chemomet.* 2001;15:773-788.
12. Lira I. *Evaluating the Measurement Uncertainty.* Bristol, UK: Institute of Physics Publishing; 2002.
13. Mandel J. *The Statistical Analysis of Experimental Data.* New York, NY: Wiley; 1964.
14. Seber GAF, Wild CJ. *Nonlinear Regression.* New York, NY: Wiley; 1989.
15. Draper NR, Smith H. *Applied Regression Analysis.* 3rd Edition. New York, NY: Wiley; 1998.
16. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* New York, NY: Springer-Verlag; 2001.
17. Jackson JE. *A User's Guide to Principal Components.* New York, NY: Wiley; 1991.
18. Martens H, Naes T. *Multivariate Calibration.* Chichester, UK: Wiley; 1989.
19. Martens H, Mevik B-H. Understanding the collinearity problem in regression and discriminant analysis. *J Chemomet.* 2001;15:413-426.
20. Wentzell PD, Andrews DT, Hamilton DC, Faber K, Kowalski BR. Maximum likelihood principal component analysis. *J Chemomet.* 1997;11:339-366.
21. Wentzell PD, Andrews DT, Kowalski BR. Maximum likelihood multivariate calibration. *Anal Chem.* 1997;69:2299-2311.
22. Geladi P, Kowalski BR. Partial least-squares regression: A tutorial. *Anal Chim Acta.* 1986;185:1-17.
23. Haaland DM, Thomas EV. Partial least-squares methods for spectral analysis. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal Chem.* 1988;60:1193-1202.
24. Helland IS. Some theoretical aspects of partial least squares regression. *Chemomet Intell Lab Syst.* 2001;58:97-107.
25. Helland IS. On the structure of partial least squares regression. *Commun Stat Simulat Comput.* 1988;17:581.
26. Höskuldsson A. *Prediction Methods in Science and Technology*. Ventura, CA: Thor Publishing; 1996.
27. Wold S, Sjöström M, Eriksson L. PLS-regression: A basic tool of chemometrics. *Chemomet Intell Lab Syst.* 2001;58:109-130.
28. Kaspar MH, Ray WH. Partial least squares modelling as successive singular value decompositions. *Comput Chem Eng.* 1993;17:985-989.
29. Höskuldsson A. PLS regression methods. *J Chemomet.* 1988;2:211-228.
30. Phatak A. *Evaluation of Some Multivariate Methods and Their Applications in Chemical Engineering*. PhD Thesis. Waterloo, Ontario, Canada: Dept. of Chemical Engineering, University of Waterloo; 1993.
31. Helland IS. Rotational symmetry, model reduction and optimality of prediction from the PLS population model. Proc of 2nd Int Symp on PLS and Related Methods, Capri, Italy, October; 2001.
32. Helland IS. Partial least squares regression. *Encyclopedia of Statistical Sciences.* 2nd Edition. Hoboken, NJ: Wiley; 2002.
33. Reis MS, Saraiva PM. Heteroscedastic latent variable modelling with applications to multivariate statistical process control. Accepted for publication in *Chemomet Intell Lab Syst.*
34. Rooney WC, Biegler LT. Design for model parameter uncertainty using nonlinear confidence regions. *AIChE J.* 2001;47:1794-1804.
35. Burnham AJ, Macgregor JF, Viveros R. Latent variable multivariate regression modeling. *Chemomet Intell Lab Syst.* 1999;48:167-180.
36. MacGregor JF, Kourti T. Multivariate statistical treatment of historical data for productivity and quality improvements. Proc of Foundation of Computer Aided Process Operations (FOCAPO 98), Snowbird, UT, July 5–10; 1998.
37. Walczak B, Massart DL. Dealing with missing data. *Chemomet Intell Lab Syst.* 2001;58:Part I: 15-27, Part II: 29-42.
38. Nelson PRC, Taylor PA, MacGregor JF. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemomet Intell Lab Syst.* 1996;35:45-65.
39. Carvalho MGV, Martins AA, Figueiredo MML, Kraft pulping of Portuguese Eucalyptus Globulus: effect of process conditions on yield and pulp properties. *Appita.* 2003. 267.
40. Bernardo FP, Pistikopoulos EN, Saraiva PM. Integration and computational issues in stochastic design and planning optimization problems. *Ind Eng Chem Res.* 1999;38:3056-3068.
41. Ku W, Storer RH, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis. *Chemomet Intell Lab Syst.* 1995;30:179-196.
42. Ricker NL. The use of biased least-squares estimators for parameters in discrete-time pulse-response models. *Ind Eng Chem Res.* 1988;27: 343-350.
43. Shi R, MacGregor JF. Modeling of dynamic systems using latent variable and subspace methods. *J Chemomet.* 2000;14:423-439.
44. de Jong S, Wise BW, Ricker NL. Canonical partial least squares and continuum power regression. *J Chemomet.* 2001;15:85-100.

Manuscript received May 25, 2004, and revision received Mar. 7, 2005.